

# Hacia un método computacional para detectar léxico contrastivo

Damián Eliel Aleman

27 de noviembre de 2017

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
  - Twitter
  - Búsquedas geolocalizadas
  - Tokenización y normalización
- 4 Métricas para detectar léxico contrastivo
  - Primeras métricas:  $\text{MaxDif}$  y  $\text{MaxDif}_g$
  - Entropía y valor de la información
  - Valores contrastivos
- 5 Análisis de las palabras contrastivas encontradas
  - Caracterización de las palabras identificadas como contrastivas
  - Validación estadística
- 6 Conclusiones y trabajo futuro

# Qué es una palabra contrastiva

Se dice que una palabra es *contrastiva* cuando la frecuencia de uso en distintas regiones es muy diferente.

## Ejemplos palabras contrastivas Argentina - España

- “che”
- “metegol”

## Ejemplos palabras contrastivas dentro de Argentina

- “gurisada”

¿Para qué sirve conocer las palabras contrastivas?

Palabra:Manso

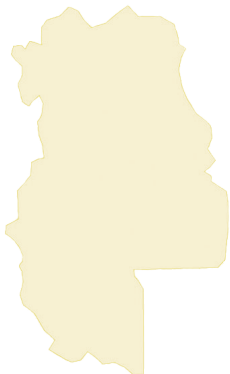


Figura: Frecuencia de uso en Mendoza

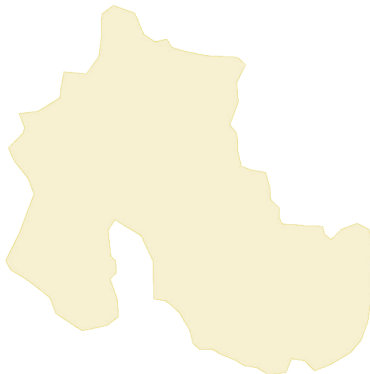


Figura: Frecuencia de uso en Jujuy

# Motivación de un método computacional para detectar léxico contrastivo

¿Cómo se conocían las palabras contrastivas? **Mediante encuestas.**

- Costosas de realizar
- Muy difícil de hacer de forma balanceada en distintas regiones de un país o de un continente.
- Más difícil aún es encuestar a una gran cantidad de personas.
- Se basan en el conocimiento *a priori*

# Temario

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
  - Twitter
  - Búsquedas geolocalizadas
  - Tokenización y normalización
- 4 Métricas para detectar léxico contrastivo
  - Primeras métricas:  $\text{MaxDif}$  y  $\text{MaxDif}_g$
  - Entropía y valor de la información
  - Valores contrastivos
- 5 Análisis de las palabras contrastivas encontradas
  - Caracterización de las palabras identificadas como contrastivas
  - Validación estadística
- 6 Conclusiones y trabajo futuro

# Web como un corpus

Kilgariff comentaba la riqueza de la Web para acceder a un volúmen de información, antes impensado.

## Ventajas

- Gratuito
- Gran cantidad de datos
- Disponible e inmediato

## Utilidades

- Corrección de ortografía
- Traducción de frases
- Estimar el tamaño de la Web

## Cita

“La web es un corpus sucio, pero el uso esperado es mucho más frecuente que lo que puede considerarse como ruido.” [KG03, p. 342]

# Temario

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
  - Twitter
  - Búsquedas geolocalizadas
  - Tokenización y normalización
- 4 Métricas para detectar léxico contrastivo
  - Primeras métricas:  $\text{MaxDif}$  y  $\text{MaxDif}_g$
  - Entropía y valor de la información
  - Valores contrastivos
- 5 Análisis de las palabras contrastivas encontradas
  - Caracterización de las palabras identificadas como contrastivas
  - Validación estadística
- 6 Conclusiones y trabajo futuro



## Qué es Twitter

- Red social creada en el 2006
- Usuarios variados: personas, instituciones gubernamentales, ONG, etc.
- Cada usuario puede escribir tuits con una longitud máxima de 140 caracteres<sup>a</sup>
- **Todos los tuits son públicos**

---

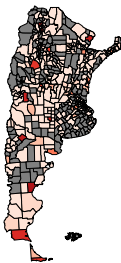
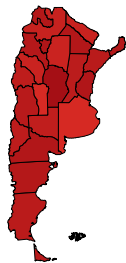
<sup>a</sup>Recientemente han aumentado el límite a 280 caracteres.

## Porqué Twitter

- Provee una interfaz pública para obtener tuits de cualquier persona
- Tópicos muy variados (gran diferencia con portales de noticias, por ejemplo)
- Escalabilidad

# Búsquedas geolocalizadas

La búsqueda geolocalizada es una herramienta que nos da la posibilidad de obtener tuits generados en un área geográfica particular.



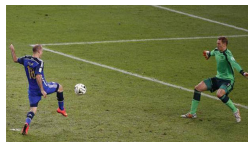
- Se realizaron búsquedas en todos los departamentos de las provincias argentinas.
- Nos quedamos con los usuarios que tienen como campo *location* al menos uno de los nombres de las ciudades de la provincia.

# Distribución temporal de tuits

¿Y si en un momento dado la mayoría de los usuarios hablan del mismo tema por un fenómeno particular?

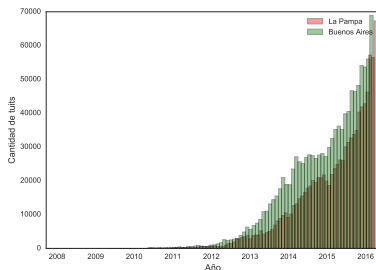
# Distribución temporal de tuits

¿Y si en un momento dado la mayoría de los usuarios hablan del mismo tema por un fenómeno particular?

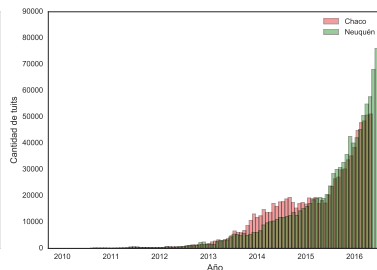


# Distribución temporal de tuits

¿Y si en un momento dado la mayoría de los usuarios hablan del mismo tema por un fenómeno particular?



(a)



(b)

**Figura:** (a) Histograma de la cantidad de tuits que se hicieron por intervalo de tiempo en la provincias La Pampa y Buenos Aires. (b) Gráfico para Chaco y Neuquén.

# Balanceo de cantidad de tuits y de usuarios

Provincia	#Palabras Distintas	#Usuarios	#Tuits	#Total Palabras
Buenos Aires	191919	920	1125042	8974372
Catamarca	173104	957	1057019	8161309
Chaco	169476	964	976943	7605991
Chubut	182592	954	1023373	8884745
Córdoba	207307	987	1224266	10075932
Corrientes	183292	939	1044951	8426940
Entre Ríos	188679	969	1193693	9462986
Formosa	169254	903	923352	7184382
Jujuy	171064	971	678004	5951778
La Pampa	186593	935	1085757	8996318
La Rioja	186041	946	704044	6757277
Mendoza	193708	945	1099717	9402399
Misiones	168400	972	984218	7790197
Neuquén	188038	927	1111201	9021449
Río Negro	194383	965	1215361	9991831
Salta	188402	884	830916	7506652
San Juan	183546	926	1002322	8377792
San Luis	164185	896	1006464	8327093
Santa Cruz	174089	935	876621	7432923
Santa Fe	201879	937	1019620	8862328
S. del Estero	166540	887	944109	7355729
T. del Fuego	197273	964	976426	8559218
Tucumán	195643	962	1093874	9238526

Cuadro: Cantidades del conjunto de datos

# ¿Qué es una palabra?

## Tokenización

Se consideró una palabra a las secuencias de caracteres formados únicamente por letras. Por lo tanto se eliminaron las menciones con @, los hashtags y links entre otros. Decidimos ignorar estos términos ya que no tienen interés lingüístico y agregarían mucho ruido a los datos.

## Normalización

Todas las letras se convirtieron a letra minúscula y las palabras con más de tres letras iguales de forma consecutiva se redujeron para que solo tengan tres repeticiones. De esta forma, el término *padreeeee* y *padreeee* fueron reducidos a una única unidad léxica (*padreee*). Esto se hizo con la librería *TweetTokenizer* de *NLTK*.

# Temario

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
  - Twitter
  - Búsquedas geolocalizadas
  - Tokenización y normalización
- 4 Métricas para detectar léxico contrastivo
  - Primeras métricas:  $\text{MaxDif}$  y  $\text{MaxDif}_g$
  - Entropía y valor de la información
  - Valores contrastivos
- 5 Análisis de las palabras contrastivas encontradas
  - Caracterización de las palabras identificadas como contrastivas
  - Validación estadística
- 6 Conclusiones y trabajo futuro



## Primeras métricas: MaxDif

Para cada palabra  $\omega$  y cada par de provincias  $p_1$  y  $p_2$ , podemos calcular el cociente entre la frecuencia máxima de  $\omega$  en ambas provincias y la frecuencia mínima:

$$\text{maxDif}(\omega, p_1, p_2) = \frac{f_{\max}(\omega, p_1, p_2)}{f_{\min}(\omega, p_1, p_2)} \quad (1)$$

Desventajas:

- 1 Un valor para cada par de provincias.
- 2 No se considera la dispersión de los valores en todas las provincias.

## Primeras métricas: $MaxDif_g$

Considerando las frecuencias de una palabra  $\omega$  sobre todas las provincias, definimos:

$$maxDif_g(\omega) = \frac{f'_{max}(\omega)}{f'_{min}(\omega)} \quad (2)$$

donde  $f'_{max}(\omega)$  es la frecuencia máxima de la palabra  $\omega$  entre las frecuencias de todas las provincias y  $f'_{min}(\omega)$  es la frecuencia mínima distinta de 0.

Con  $maxDif_g$  se resume en un único valor la contrastividad de la palabra, sin embargo sigue sin considerar la distribución de las frecuencias.

# La entropía de la información

Sea  $\mathbf{p} = (p_1, \dots, p_n)$  un vector de probabilidad puntual. Es decir,  $p_i \geq 0$  y  $\sum_{i=1}^n p_i = 1$ .

Definimos la entropía de  $\mathbf{p}$  siendo

$$H(\mathbf{p}) = - \sum_{i=1}^n \log(p_i) p_i. \quad (3)$$

# La entropía de la información

Sea  $\mathbf{p} = (p_1, \dots, p_n)$  un vector de probabilidad puntual. Es decir,  $p_i \geq 0$  y  $\sum_{i=1}^n p_i = 1$ .

Definimos la entropía de  $\mathbf{p}$  siendo

$$H(\mathbf{p}) = - \sum_{i=1}^n \log(p_i) p_i. \quad (3)$$

Dime que tan uniforme eres y te diré cuanta entropía tienes

Las palabras utilizadas más uniformemente en las distintas provincias como *de* o *que* aportan menos información que la palabra *celular*.

# La entropía de la información

Sea  $\mathbf{p} = (p_1, \dots, p_n)$  un vector de probabilidad puntual. Es decir,  $p_i \geq 0$  y  $\sum_{i=1}^n p_i = 1$ .

Definimos la entropía de  $\mathbf{p}$  siendo

$$H(\mathbf{p}) = - \sum_{i=1}^n \log(p_i) p_i. \quad (3)$$

## Observaciones

- 1  $H(\mathbf{p}) = 0$  si y solo si  $\mathbf{p}$  esta concentrada en un unico punto: existe  $i$  tal que  $p_i = 1$  y  $p_j = 0$ , para todo  $j \neq i$ .
- 2 La función  $H$  se maximiza tomando  $\mathbf{p}$  equiprobable:  $p_i = 1/n$ , para todo  $i$ .

# Valor de la información

Zanette y Montemurro definieron al *valor de la información de una palabra* como

$$\Delta I_w(\omega) = p(\omega) (\hat{H}(\omega) - H(\omega)) = p(w) \Delta H(\omega) \quad (4)$$

siendo  $p(\omega)$  la frecuencia total de la palabra en el texto.

# Valor contrastivo sobre las palabras

## Valor contrastivo sobre las palabras

$$I_w(\omega) = \text{norm}_w(\omega) \cdot (\hat{H}_w(\omega) - H_w(\omega)) \quad (5)$$

donde  $\text{norm}_w$  sirve para normalizar sobre la cantidad de ocurrencias de la palabra.

## Observaciones

- Si una palabra se dice muchas veces el valor de  $I_w(\omega)$  es más alto.
- Si dos palabras se dicen la misma cantidad de veces, la palabra que tenga una dispersión más heterogenea será la de mayor valor contrastivo sobre las palabras.

# Robusteciendo métrica: valor contrastivo sobre las personas

¿Y si algunas palabras tienen un  $I_w$  alto debido a pocas personas que las mencionan constantemente?

Hay que tener en cuenta a la distribución de la cantidad de personas que mencionan cada palabra en las provincias.

## Valor contrastivo sobre las personas

$$I_p(\omega) = \text{norm}_p(\omega) \cdot (\hat{H}_p(\omega) - H_p(\omega)) \quad (6)$$

donde  $\text{norm}_p$  sirve para normalizar sobre la cantidad de personas que mencionan la palabra  $\omega$ .

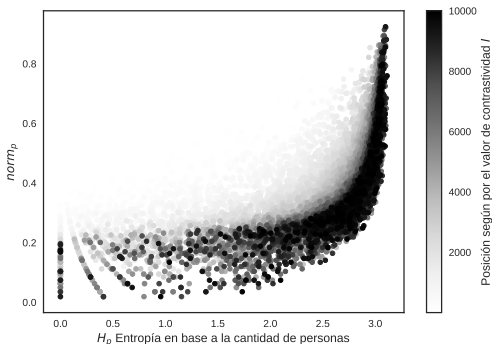


# Robusteciendo métrica: valor contrastivo sobre las personas

## Valor contrastivo sobre las personas

$$I_p(\omega) = \text{norm}_p(\omega) \cdot (\hat{H}_p(\omega) - H_p(\omega)) \quad (6)$$

donde  $\text{norm}_p$  sirve para normalizar sobre la cantidad de personas que mencionan la palabra  $\omega$ .

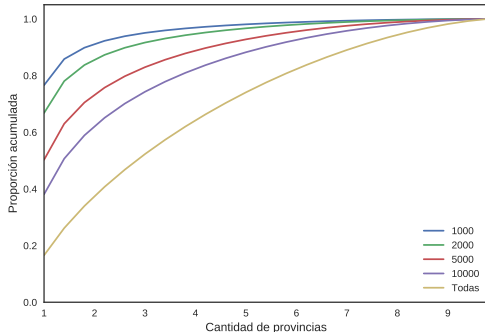


# Valor de contrastividad

## Valor de contrastividad

Como nos interesa tanto la distribución de la cantidad de ocurrencias de cada palabra, como la de la cantidad de usuarios que la menciona, definimos

$$I(\omega) = I_w(\omega) \cdot I_p(\omega) \quad (7)$$

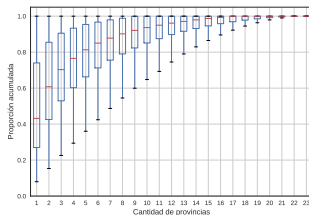


Proporción de ocurrencias acumulada según la muestra de palabras.

# Proporción de ocurrencias

Detectadas las palabras que se utilizan en forma diferenciada, interesa detectar donde se utilizan mas.

- 1 por cada subconjunto de palabras:
- 2 por cada palabra  $w$ :
- 3 se ordenen las provincias en funcion de la cantidad de veces que se utiliza  $w$
- 4 por  $k = 1..23$ :
- 5 se calcula la proporcion acumulada de ocurrencias por las primeras  $k$  provincias

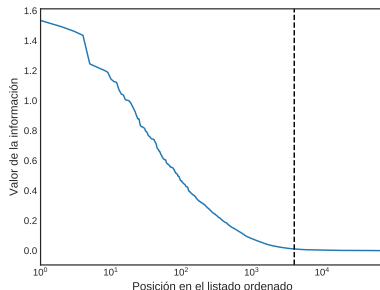


# Temario

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
  - Twitter
  - Búsquedas geolocalizadas
  - Tokenización y normalización
- 4 Métricas para detectar léxico contrastivo
  - Primeras métricas:  $\text{MaxDif}$  y  $\text{MaxDif}_g$
  - Entropía y valor de la información
  - Valores contrastivos
- 5 Análisis de las palabras contrastivas encontradas
  - Caracterización de las palabras identificadas como contrastivas
  - Validación estadística
- 6 Conclusiones y trabajo futuro

# Palabras candidatas

Para buscar las palabras candidatas a tener contrastes significativos en cuanto a la cantidad de ocurrencias en distintas provincias, elegimos el conjunto de las primeras cinco mil (5000) palabras con mayor valor de nuestra métrica.



# Pipeline



- **Coloquialismos o vulgarismos**

“Perdon pero tenes que ser muy **culiado/a** para ir a mc y pedirte una ensalada” (Córdoba)

“Q **chombi** hacer un chiste y q la otra persona no se ría o no lo entienda” (Mendoza)

- **Indigenismos**

“Te regalo ser **mitaí** y ir a jurar la bandera con el guardapolvo caliente ese y la corbata que te ahorca todo (Del guaraní mitaí “pequeño”)” (Formosa)

“**Angá** mi negrito, esta triste (Del guaraní angá aprox. “pobre”)” (Corrientes)

- **Gentilicios**

**Casildense** (de Casilda), **concordiense** (de Concordia) y **obereño** (de Oberá).

# Más resultados

- **Leísmo**

“No te olvides de **saludarle** a tu suegro hoy” (Misiones)

“Vine a **visitarle** a mis primas y estan re colgadas, para eso me quedaba en mi casa no maaa ” (Misiones)

“A **esperarle** a nahuel, que traiga los teresss ” (Formosa)

- **Voces sospechadas generales pero con acepción local diferente**

“Mañana que alguien **atine** con parque y porrones” (Mendoza)

“**Mansas** ganas de sentarme a tomar un te con semitas” (San Juan)

“**Habilítenme** una nueva espaldaa” (Tierra del Fuego)

“sigo **asada** por cosas que han pasado hace como dos dias, que falla (Mendoza) / Que **asada** estoy, tengo la cabeza echa un lío” (San Juan)

- **Voces con una morfología propia de una región**

Ejemplo: terminación azo/aza con base adjetiva.

“Creo que va a estar **malazo** lo de esta noche ” (San Juan)

“Esta **locaza** esa mina para hacer eso” (Córdoba)



# Problema desde el punto de vista estadístico

## Hipótesis nula: $H_0$

La palabra tienen un uso homogéneo en las distintas regiones de la Argentina, es decir que la frecuencia de ocurrencias de cada palabra debería ser similar independientemente de la región.

# Test hipergeométrico

Para aplicar el test hipergeométrico representamos los datos sobre la palabra en una tabla de 2x2 como la de la siguiente Tabla.

	#Palabras sobre región	#Palabras en el resto de Argentina	Total
# Palabras $w$	$k$	$K - k$	$K$
# Palabras $\neq w$	$n - k$	$N + k - n - K$	$N - K$
Total	$n$	$N - n$	$N$

# Test t de Welch

El test de Welch nos provee un valor de probabilidad para rechazar la hipótesis nula que afirma que las medias de las dos distribuciones son iguales.

# Test t de Welch

El test de Welch nos provee un valor de probabilidad para rechazar la hipótesis nula que afirma que las medias de las dos distribuciones son iguales.

Las suposiciones del test

- 1 Todos los textos son estadísticamente independientes
- 2 La media de las frecuencias proviene de una distribución normal

# Test t de Welch

El test de Welch nos provee un valor de probabilidad para rechazar la hipótesis nula que afirma que las medias de las dos distribuciones son iguales.

## Metodología

Agrupamos todos los tuits de cada usuario representando un texto.<sup>a</sup>

**Corpus S** Todos los textos de los usuarios que provienen de las provincias en donde se cubre el 80 % de las ocurrencias

**Corpus T** Los textos creados por usuarios del resto de las provincias

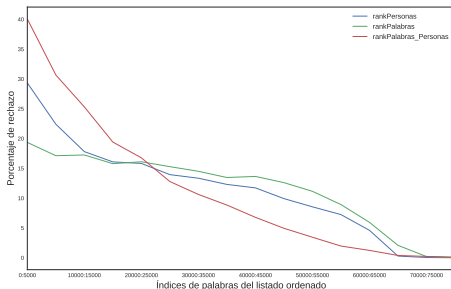
---

<sup>a</sup>Notar que la suposición de independencia es más débil.

# Resultados test t de Welch

- 1 Para cada métrica  $I$ ,  $I_W$ ,  $I_P$  variamos los subconjuntos de palabras de acuerdo al listado ordenado según estas.
- 2 Calculamos la tasa de rechazo de la hipótesis nula, definida por:

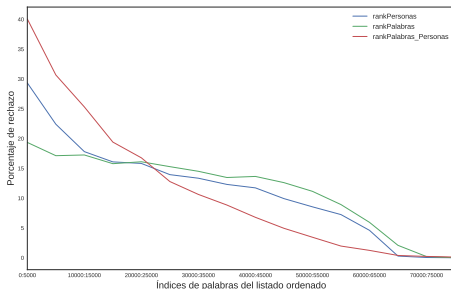
$$\text{Tasa de rechazo}(\text{tests}) = \frac{\#\{t : \text{tests} \mid p - \text{valor}(t) < 0,05\}}{\#\text{tests}} \quad (8)$$



# Resultados test t de Welch

- 1 Para cada métrica  $I$ ,  $I_W$ ,  $I_P$  variamos los subconjuntos de palabras de acuerdo al listado ordenado según estas.
- 2 Calculamos la tasa de rechazo de la hipótesis nula, definida por:

$$\text{Tasa de rechazo}(\text{tests}) = \frac{\#\{t : \text{tests} \mid p - \text{valor}(t) < 0,05\}}{\#\text{tests}} \quad (8)$$



# Temario

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
  - Twitter
  - Búsquedas geolocalizadas
  - Tokenización y normalización
- 4 Métricas para detectar léxico contrastivo
  - Primeras métricas:  $\text{MaxDif}$  y  $\text{MaxDif}_g$
  - Entropía y valor de la información
  - Valores contrastivos
- 5 Análisis de las palabras contrastivas encontradas
  - Caracterización de las palabras identificadas como contrastivas
  - Validación estadística
- 6 Conclusiones y trabajo futuro



# Conclusiones

- Desarrollamos una métrica de la contrastividad de uso de una palabra en distintas regiones.
- Para probar esta métrica recolectamos un conjunto de datos de textos de la Argentina a través de la API de Twitter.
- Obtuvimos aproximadamente 1 palabra contrastiva relevante lingüísticamente cada 17 palabras.
- Varias de las palabras detectadas a partir de la métrica desarrollada serán agregadas al Diccionario del habla de los argentinos.

# Trabajo a futuro

- Reproducir el trabajo para todos los países hispanoparlantes.
- Obtener regiones dialectales a partir de métodos de clustering, lo cual permitiría validar la vigencia de las regiones propuestas por Vidal de Battini en 1964 [VdB64].
- Analizar la contrastividad léxica comparando la distribución de n-gramas.

# ¿Preguntas?



Adam Kilgarriff and Gregory Grefenstette.

Introduction to the special issue on the web as corpus.

*Computational linguistics*, 29(3):333–347, 2003.



Berta Elena Vidal de Battini.

El español en la argentina.

Technical report, Argentina., 1964.